Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 493-515

Exploring rater errors and systematic biases using adjacent-categories Mokken models

Stefanie A. Wind[1] *& George Engelhard, Jr.*[2]

Abstract

Adjacent-categories formulations of polytomous Mokken Scale Analysis (ac-MSA) offer insight into rating quality in the context of educational performance assessments, including information regarding individual raters' use of rating scale categories and the degree to which student performances are ordered in the same way across raters. However, the degree to which ac-MSA indicators of rating quality correspond to specific types of rater errors and systematic biases, such as severity/leniency and response sets, has not been fully explored. The purpose of this study is to explore the degree to which ac-MSA provides diagnostic information related to rater errors and systematic biases in the context of educational performance assessments. Data from a rater-mediated writing assessment are used to explore the sensitivity of ac-MSA indices to two categories of rater errors and systematic biases: (1) rater leniency/severity; and (2) response sets (e.g., centrality). Implications are discussed in terms of research and practice related to large-scale educational performance assessments.

Keywords: Mokken scaling; rater errors; Rasch measurement theory

[1] *Correspondence concerning this article should be addressed to:* Stefanie A. Wind, PhD, Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, 313C Carmichael Hall, USA; email: swind@ua.edu

[2] The University of Georgia, USA

Concerns related to rating quality are prevalent in research on educational performance assessments (Hamp-Lyons, 2007; Lane & Stone, 2006; Saal, Downey, & Lahey, 1980). Accordingly, researchers have proposed numerous quantitative techniques that can be used to evaluate ratings, including indicators of rater errors and systematic biases. These techniques for evaluating rating quality reflect a variety of measurement frameworks, including methods based on observed ratings (i.e., Classical Test Theory) and methods based on scaled ratings (i.e., Item Response Theory; IRT). Whereas most methods based on observed ratings provide group-level indicators of rating quality, such as proportions of rater agreement or reliability coefficients, methods based on scaled ratings provide information about rating quality at the individual rater level (Wind & Peterson, 2017).

When ratings are evaluated with the purpose of improving the measurement quality of an assessment system, such as during rater training or monitoring procedures, diagnostic information is needed that describes rating quality at the individual rater level. In particular, information that describes the degree to which individual raters demonstrate specific types of rater errors and systematic biases, such as severity/leniency and central tendency, can provide useful feedback for improving rating quality that goes beyond overall summaries of rater agreement or reliability (Myford & Wolfe, 2003, 2004; Wolfe & McVay, 2012).

Currently, most research on quantitative rating quality indices is based on parametric IRT models. IRT models are classified as parametric when they involve the transformation of ordinal ratings to an interval-level scale. The practical implications of this transformation are that the *rater response function*, or the relationship between student achievement estimates and ratings is assumed to conform to a particular shape (usually the logistic ogive) that matches a specific distribution, and that the measures of student achievement and rater severity are estimated on an interval-level scale. It is also possible to examine rating quality using nonparametric IRT methods, which do not involve the transformation between ordinal ratings and an interval-level scale. In particular, Mokken Scale Analysis (MSA; Mokken, 1971) can be applied to data from rater-mediated educational assessments in order to evaluate rating quality (Snijders, 2001; Wind & Engelhard, 2015). The MSA approach is based on less-strict underlying requirements for ordinal ratings compared to parametric IRT models (Meijer, Sijtsma, & Smid, 1990). Although they are less strict, MSA is nonetheless characterized by underlying requirements for ratings that can be used to evaluate the degree to which raters demonstrate basic measurement properties, without the requirement of a parametric model (Wind & Engelhard, 2015). The MSA approach to evaluating rating quality provides an exploratory approach to examining the psychometric properties of ratings that can be used to examine the measurement properties associated within individual raters, prior to the application of a parametric model.

Recently, a nonparametric procedure based on Mokken Scale Analysis (MSA) was presented that can be used to explore rating quality at the individual rater level (Wind, 2016; Wind & Engelhard, 2015). This model is essentially an application of the Nonparametric Partial Credit Model (Hemker, Sijtsma, Molenaar, & Junker, 1997) to the context of rater-mediated assessments that also includes indices of psychometric properties based on MSA, such as monotonicity, double monotonicity, and scalability coefficients. Relatedly, ac-MSA can be described as a special case of Mokken's (1971) Monotone Homogeneity model (Van Der Ark, 2001). Because it is nonparametric, this approach can be used to evaluate individual raters in terms of fundamental measurement properties, including scalability, monotonicity, and invariance in rater-mediated assessments without imposing potentially inappropriate transformations on the ordinal rating scale. In particular, the application of adjacent-categories MSA models (ac-MSA; Wind, 2016) to rater-mediated assessments has been shown to offer valuable insight into rating quality, including information regarding individual raters' use of rating scale categories and the degree to which student performances are ordered in the same way across raters. However, the degree to which rating quality indicators based on ac-MSA correspond to specific types of rater errors and systematic biases has not been fully explored.

Purpose

The purpose of this study is to explore the degree to which ac-MSA provides diagnostic information related to rater errors and systematic biases in the context of educational performance assessments. Specifically, this study focuses on the use of numeric and graphical indicators based on ac-MSA to identify two major categories of rater errors and systematic biases: (a) leniency/severity; and (b) response sets (e.g., centrality. Two research questions guide the analyses:

1. How can ac-MSA be used to detect rater leniency/severity?
2. How can ac-MSA be used to detect rater response sets?

In order to provide a frame of reference for exploring rater errors and systematic biases using ac-MSA, indicators of rater leniency/severity and response sets are calculated using the Rasch Partial Credit (PC) model (Masters, 1982). Then, indicators of rating quality based on ac-MSA are explored as they relate to rater classifications based on the Rating Scale (RS) model (described further below).

Rater errors and systematic biases

As noted above, researchers have proposed numerous quantitative techniques for evaluating the quality of ratings in performance assessments. Although there are some discrepancies in the terminology and methods used to calculate these indices (Saal et al., 1980), most rating quality indices reflect similar concerns. In particular, rating quality indicators that are used in practice generally reflect concerns related to the degree to which raters assign the same or similar scores to the same student performances (rater agreement), or the degree to which raters consistently rank-order student performances (rater reliability; Johnson, Penny, & Gordon, 2009; Wind & Peterson, 2017).

In addition to rater agreement and reliability, indicators of rating quality can also be used to identify specific types of rater errors and systematic biases that can lead to targeted

Table 1:
Definitions of Rater Errors and Systematic Biases

Rater Errors and Systematic Biases	Definition
Rater Severity/Leniency	Raters systematically assign lower-than-expected ratings (severity) or higher-than-expected ratings (leniency) than is warranted by the quality of student performances
Response Sets	Raters assign rating patterns that suggest the idiosyncratic interpretation and use of rating scale categories, such as centrality, and muted/noisy ratings.

rater remediation or the revision of scoring materials, such as rubrics, score-level exemplars, and performance level descriptors (Engelhard, 2002; Wolfe & McVay, 2012). Although researchers have described many different types of errors and systematic biases, two major categories have been particularly useful for classifying rating patterns that may warrant further attention in the context of educational performance assessments: (a) severity/leniency; and (b) response sets. Table 1 includes definitions for these two categories of rater errors and systematic biases, and these definitions are elaborated and illustrated below.

Severity/Leniency

The first major category of rater errors and systematic biases is *rater severity/leniency*. In general, raters are considered severe or lenient when they systematically assign lower-than-expected or higher-than-expected ratings, respectively, than is warranted by the quality of student performances. Table 2, Panel A includes a small illustration that illustrates rater severity/leniency for ten student performances based on a rating scale with five categories (1=*low*, 5=*high*). The illustration includes a criterion rater whose ratings reflect “known” or “true” scores, a severe rater, and a lenient rater. In the illustration, the severe rater consistently assigns lower ratings than the criterion rater, and the lenient rater consistently assigns higher ratings than the criterion rater.

Response sets

The second major category of rater errors and systematic biases is *response sets*. Rater response sets include a variety of rating patterns that suggest the idiosyncratic interpretation and use of rating scale categories. Researchers have described numerous types of response sets that are viewed as potentially problematic in rater-mediated assessments. Among these response sets, a common classification includes *range restriction*, which is the tendency for raters to use only a subset of the rating categories when performances

Table 2:
Illustrations of Rater Errors and Systematic Biases

Panel A: Severity/Leniency										
Raters	Performances									
	1	2	3	4	5	6	7	8	9	10
Criterion	*5*	*2*	*1*	*4*	*3*	*1*	*4*	*3*	*1*	*5*
Severe	3	1	1	2	1	1	2	1	1	3
Lenient	5	3	2	5	4	3	4	5	2	5
Panel B: Response Sets										
Raters	Performances									
	1	2	3	4	5	6	7	8	9	10
Criterion	*5*	*2*	*1*	*4*	*3*	*1*	*4*	*3*	*1*	*5*
Central	3	3	3	4	2	3	3	3	3	4
Muted	4	3	3	4	4	3	4	4	3	4
Noisy	2	1	4	2	5	5	1	1	4	2

warrant ratings across the range of the scale. Although range restriction can occur anywhere along the rating scale, a common form of range restriction is rater *centrality* (i.e., central tendency), which occurs when raters use the middle categories more frequently than expected. Continuing the small illustration described above, Table 2, Panel B illustrates rater centrality using the same criterion rater from Panel A and a central rater. Whereas the criterion rater uses the full range of rating scale categories, the central rater consistently assigns scores in the central categories of the rating scale.

When Rasch models are used to explore rating quality, idiosyncratic rating patterns are frequently described as *muted* or *noisy.* Specifically, rating patterns are described as muted when there is less variation than expected by the model (e.g., in the case of range restriction), and noisy when there is more variation than expected by the model. For example, a muted rating pattern might match the example response sets described above for rater centrality, another type of range restriction, such as the muted pattern illustrated in Table 2 Panel B. The example muted rater consistently assigns ratings in categories 3 or 4. The illustration also includes a noisy rating pattern, where the example rater's responses include seemingly random noise, or ratings that appear more haphazard than those expected based on student performances.

Methods

In order to explore the research questions for this study, we used ac-MSA (described further below) to explore data from a rater-mediated writing assessment in terms of rater severity/leniency and response sets. Because indicators of these rater errors and systematic biases based on Rasch measurement theory are more well known in the psychometric literature (e.g., Eckes, 2015; Engelhard, 2002; Myford & Wolfe, 2003, 2004), indicators based on the Rasch PC model are used as a frame of reference for interpreting Mokken indices within these categories. This section includes a description of the instrument and methods for exploring the rating data.

Instrument

Data were collected during a recent administration of the Alaska High School Writing (AHSW) test. The subset of ratings included in the current sample includes 40 raters who scored essays composed by 410 students using a five-category rating scale. All of the raters scored all 410 students, such that the rating design was fully crossed (Engelhard, 1997). For illustrative purposes, we focus on ratings of student responses to one of the essay prompts in the current analysis.

Procedures

Our data analysis procedures included two major steps. First, rating quality indicators based on the PC model were calculated and used to classify each of the 40 raters in terms of severity/leniency and response sets. Second, indicators of measurement quality based on ac-MSA were explored within the rater classifications based on the PC model. We conducted the PC model analyses using Facets (Linacre, 2015), and we conducted the ac-MSA analyses using R (R Core Team, 2015); code for both approaches is available from the first author upon request.

Rasch rating quality indicators

As noted above, numerous scholars have explored rating quality in performance assessments using indicators of measurement quality based on Rasch models for polytomous ratings, including the Rasch rating scale (RS) model (Andrich, 1978), the Rasch partial credit (PC) model (Masters, 1982), and RS and PC formulations of the Many-Facet Rasch (MFR) model (Linacre, 1989). The PC formulation of the polytomous Rasch model was selected for this study because it facilitates examination of rating scale category use more explicitly than does the RS version of the model. This model is stated mathematically as follows:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \lambda_i - \tau_{ik}\ , \tag{1}$$

where

P_{nik} = the probability of Student *n* receiving a rating in category *k* from Rater *i*,

P_{nik-1} = the probability of Student *n* receiving a rating in category *k*–1 from Rater *i*,

θ_n = the location of Student *n* on the construct (i.e., ability),

λ_i = the location of Rater *i* on the construct (i.e., severity), and

τ_{ik} = the location on the construct where the probability for a rating in Category *k* and *k*–1 is equally probable for Rater *i.*

When the PC model is applied to rating data, several indices can be calculated that provide diagnostic information related to the three categories of rater errors and systematic biases described above. First, rater locations on the logit scale (λ) are used as indicators of rater severity/leniency. Specifically, when the PC model in Equation 1 is estimated, the rater facet is centered at (i.e., fixed to) zero logits, such that more-severe raters have positive logit scale calibrations, and more-lenient raters have negative logit scale calibrations. Following Wolfe and McVay (2012), the critical value of +/- 0.50 logits is used to identify severe and lenient raters, such that raters with locations higher or lower than 0.50 logits from the mean rater location are considered severe or lenient, respectively.

Second, model-data fit statistics for the rater facet are used as indicators of rater response sets. Following Engelhard and Wind (in press), values of the unstandardized Outfit statistic (Outfit *MSE*) were considered for each rater. Values of Outfit *MSE* greater than +1.50 were used to identify noisy raters, and values of Outfit *MSE* less than 0.50 were used to identify muted raters.

Mokken Scale Analysis

Mokken Scale Analysis (MSA; Mokken, 1971) is a nonparametric approach to item response theory that is theoretically aligned with Rasch measurement theory. Mokken proposed an approach to evaluating the psychometric properties of social science measures that allows researchers to evaluate the requirements for invariant measurement while maintaining the ordinal level of measurement that characterizes the raw scores. Specifically, MSA provides an exploratory approach to evaluating the degree to which persons are ordered consistently across items, and items are ordered consistently across persons. As a result, researchers can use this nonparametric approach to explore fundamental measurement properties without potentially inappropriate parametric transformations or assumptions.

Dichotomous Mokken models

In the original presentation of MSA, Mokken proposed two models: (1) the Monotone Homogeneity (MH) model; and (2) the Double Monotonicity (DM) model. The MH model is based on three requirements: (1) *Monotonicity*: As student locations on the latent variable increase, the probability for correct response (*X*=1) does not decrease; (2) *Unidimensionality*: Students' responses reflect one latent variable; and (3) *Local Independence*: Students' responses to each item are not dependent on their responses to any other item, after controlling for the latent variable. In practice, adherence to the MH model is evaluated using graphical and numeric analyses, where evidence of non-decreasing item response functions (IRFs) across increasing levels of the latent variable suggest that monotonicity is observed. Scalability coefficients are also used to evaluate the MH model. These coefficients provide an index of the degree to which individual items, pairs of items, or sets of items are associated with *Guttman errors*, or the combination of a correct response to a more-difficult item in combination with an incorrect response to an easier item. Evidence of adherence to the MH model suggests that person ordering on the latent variable is invariant across items.

The DM model shares the three MH model requirements and includes a fourth requirement: (4) *Invariant item ordering*: item response functions for any given item do not intersect with response functions for any other item. In practice, adherence to the DM model is evaluated using graphical and numeric analyses, where evidence of non-intersecting IRFs suggests that double monotonicity is observed. Evidence of adherence to the DM model suggests that item ordering on the latent variable is invariant across persons.

Polytomous Mokken models

Following the original presentation of MSA, Molenaar (1982) presented polytomous versions of Mokken's original nonparametric models. Similar to polytomous extensions of other IRT models, the polytomous formulations of MSA models are based on the same requirements as the dichotomous formulations, but the requirements are evaluated at the level of rating scale categories, rather than for the overall item. Specifically, for each polytomous item with k rating scale categories, $k - 1$ Item Step Response Functions (ISRFs; τ) are calculated that reflect the difficulty associated with a rating in a particular category. In their original formulation, ISRFs for polytomous MSA models are calculated using cumulative probabilities, where each τ reflects the difficulty associated with receiving a rating in category k or any higher category, as defined based on the ordinal rating scale.

In order to extend the use of MSA to the context of educational performance assessments, Wind (2016) proposed an adaptation of polytomous MSA models, where the ISRFs are calculated using adjacent-categories probabilities. Specifically, adjacent-categories MSA (ac-MSA) models are defined such that each τ reflects the difficulty associated with receiving a rating in category k, rather than k–1. This approach is more conceptually aligned with performance assessments, where the difficulty associated with each category in the rating scale is of more interest than the difficulty associated with a cumulative set of categories (Andrich, 2015). Furthermore, the adjacent-categories formulation matches the threshold formulation that is used in polytomous Rasch models – leading to a closer theoretical alignment with polytomous Rasch models.

Mokken rating quality indices

In this study, ac-MSA models are used to explore rater errors and systematic biases. Similar to the Rasch approach to evaluating rating quality, data from rater-mediated assessments can be evaluated using ac-MSA by treating raters as a type of "item" or "assessment opportunity." Then, indices of adherence to the model requirements can be examined as evidence of rating quality.

Following the procedures that are typically used to evaluate psychometric properties based on MSA (e.g., Meijer, Tendeiro, & Wanders, 2015; Sijtsma, Meijer, & van der Ark, 2011), we focus on three indicators of rating quality. The first two indicators are based on the adjacent-categories formulation of the polytomous MH model: (A) Rater monotonicity; and (B) Rater scalability. The third indicator is based on the adjacent-categories formulation of the DM model: (C) Invariant rater ordering.

A. Rater monotonicity. Rater monotonicity refers to the degree to which the probability associated with receiving a rating in rating scale category *k,* rather than category *k*–1 increases over increasing levels of student achievement. Rater monotonicity can be evaluated for each ISRF using graphical and numeric indices. Figure 1 illustrates a graphical procedure for evaluating rater monotonicity for an example rater using a four-category rating scale. The *y*-axis shows the probability for a rating in the higher of each pair of adjacent categories [P(*X*=*k*)/P(*X*=k – 1)]. The *x*-axis shows student *restscores (R)*, which are the nonparametric analogue to person (theta) estimates in Rasch models. Restscores are calculated by subtracting the rating each student receives from the rater of interest from their total score across the rest of the raters. Then, students with the same or adjacent restscores are combined into restscore groups in order to evaluate model assumptions. The *y*-axis shows the probability for a rating in each category, rather than the category just below it. The three lines show the probability for a rating in category 1, rather than category 0 (highest line), the probability for a rating in category 2, rather than in category 1 (middle line), and the probability for a rating in category 3 rather than in category 2 (lowest line). The example monotonicity plot in Figure 1 illustrates adherence to rater monotonicity because the probability for a rating in a higher category is non-decreasing across increasing restscores.

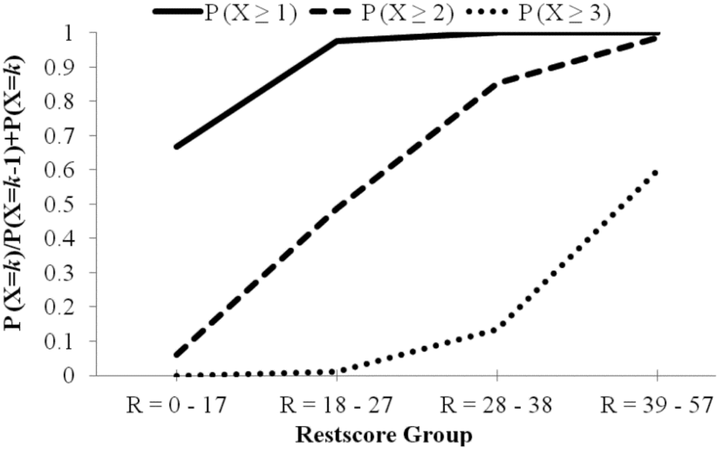


Figure 1:
Example Monotonicity Plot

Rater monotonicity can also be evaluated using statistical hypothesis tests. Specifically, for each pair of adjacent restscore groups, the following null hypothesis is evaluated: the adjacent-categories probability for a rating in a particular rating scale category is higher for the group with higher restscores than the group with lower restscores. Rejections of this null hypothesis constitute violations of rater monotonicity.

B. Rater scalability. In traditional applications of MSA, scalability coefficients are used as indicators of the degree to which individual items, pairs of items, and overall sets of items are associated with Guttman errors. Specifically, scalability coefficients are calculated using one minus the ratio of the observed and expected frequency of Guttman errors within every possible pairwise combination of items. For item pairs, the formula for the scalability coefficient is as follows:

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}} \tag{2}$$

A value of $H_{ij} = 1.00$ indicates that there are no Guttman errors associated with a particular item pair. Scalability coefficients can also be calculated for individual items (H_i) and item sets (H). For individual items, scalability coefficients (H_i) are calculated using each item pair that includes the item of interest. Scalability coefficients for item sets (H) are calculated using all of the item pairs.

Polytomous scalability coefficients can also be calculated for individual raters, pairs of raters, and a group of raters. Specifically, polytomous scalability coefficients are calculated using Guttman errors that are observed at the level of rating scale categories. Guttman errors for polytomous items are observed when the probability for a rating in a higher category is greater than the probability for a rating in a lower category. When the ac-MSA formulation is applied, Guttman errors are identified by first establishing the overall difficulty ordering of the ISRFs across items (or raters), and then identifying deviations from this ordering. When the adjacent-categories formulation is used, the difficulty ordering of ISRFs is calculated using adjacent-categories probabilities, rather than the cumulative probabilities that are usually used to calculate MSA scalability coefficients. Additional details about ac-MSA scalability coefficients can be found in Wind (2016) and Wind (under review).

In the context of rater-mediated assessments, scalability coefficients for individual raters (H_i) are diagnostically useful because they provide an index of the degree to which individual raters are associated with Guttman errors. For raters, Guttman errors suggest idiosyncratic rating patterns that warrant further investigation. Mokken (1971) suggested a minimum critical value of $H_i = 0.30$ for item selection purposes, where values between $0.30 \leq H_i < 0.40$ suggest weak scalability, values in the range of $0.40 \leq H_i < 0.50$ suggest moderate scalability, and values greater than $H_i = 0.50$ suggest strong scalability. Although these critical values are widely applied in practice (e.g., Meijer et al., 2015; Sijtsma et al., 2011), they have not been thoroughly examined in the context of polytomous items in general, as well as in the context of rater-mediated assessments more specifically.

C. Invariant rater ordering. Finally, invariant rater ordering (IRO) refers to the degree to which rater ordering in terms of severity is invariant across students. Although it is possible to evaluate invariant ordering at the level of ISRFs by examining the degree to which rating scale categories for individual raters are ordered consistently across students, most MSA researchers investigate invariant ordering at the overall item level (Ligtvoet, Van der Ark, Marvelde, & Sijtsma, 2010; Sijtsma et al., 2011). Likewise, in this study, we investigate IRO at the overall rater level using graphical displays and statistical hypothesis tests.

Figure 2 illustrates a graphical procedure for evaluating IRO for a pair of example raters using a four-category rating scale. Similar to Figure 1, the *x*-axis shows *restscores (R)*. Because the plot in Figure 1 includes two raters, restscores are calculated for each student by subtracting their ratings from the two raters of interest from their total ratings across the remaining raters. Because IRO is evaluated for overall raters, rather than within rating scale categories, the *y*-axis shows average ratings. The solid line shows the average rating assigned by Rater *i* within each restscore group, and the dashed line shows the average rating assigned by Rater *j* within each restscore group. The example raters in Figure 2 illustrate adherence to IRO because the raters are ordered consistently across all of the restscore groups, such that Rater *j* is consistently more severe than Rater *i* regardless of students' achievement level.

Similar to rater monotonicity, IRO can also be evaluated using statistical hypothesis tests. Specifically, given raters *i* and *j* who are ordered in terms of severity such that rater *i* is more lenient than rater *j* ($i < j$), the following null hypothesis is evaluated within each restscore group: the average rating from rater *i* is greater than or equal to the average rating from rater *j*. Rejections of this null hypothesis constitute violations of IRO.

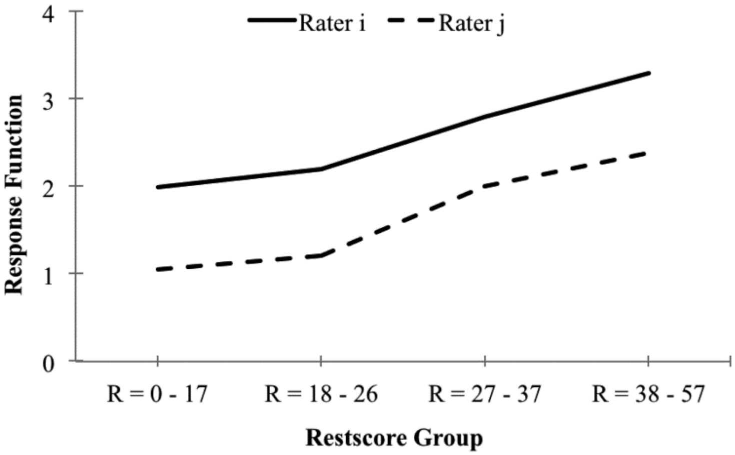


Figure 2:
Example Invariant Rater Ordering Plot

In order to explore the degree to which indicators of rating quality based on rater monotonicity, scalability, and invariant ordering can detect specific types of rater errors and systematic biases, the three categories of ac-MSA indices discussed above are applied to the entire set of ratings. Then, the prevalence of violations of monotonicity, scalability, and invariant ordering is considered within each group of raters (fair, lenient, severe, muted, and noisy).

Results

Rater classifications

Table 3 includes PC model results for each of the 40 raters, along with corresponding classifications related to rater severity/leniency and muted/noisy response sets. Overall, these results suggest that several raters who scored the AHSW test demonstrated rater errors and systematic biases that may warrant further investigation. Specifically, 11 raters were lenient ($\lambda \leq$- 0.50), eight raters were severe ($\lambda \leq$+0.50), two raters were muted (Outfit *MSE* $\leq$ 0.50), and four raters were noisy (Outfit *MSE* $\geq$ +1.50). There were 20 raters who were not classified as severe, lenient, noisy, or muted; these raters are described as “fair” in the remainder of the manuscript.

Mokken rating quality indices

Table 3 also includes results from the ac-MSA analysis for each rater; and the ac-MSA results are summarized within each of the Rasch classifications in Table 4. In this section, the ac-MSA results are described as they relate to rater monotonicity, rater scalability, and IRO.

A. Rater monotonicity

The results in Table 3 suggest that there were very few significant violations of rater monotonicity among the 40 raters who scored the AHSW test. This finding suggests that, in general, the students were ordered consistently across raters, such that the interpretation of individual students’ writing achievement was invariant across raters.

Table 4 summarizes the monotonicity results within the rater groups based on the Rasch model classifications. Specifically, within each group of raters, the average number of significant violations of rater monotonicity is presented. Across these rater classifications, it is interesting to note that violations of monotonicity were observed most frequently among raters who were classified as noisy (*M*=0.25, *SD*=0.50), and least frequently within the severe and muted rater groups (*M*=0.00, *SD*=0.00). This finding suggests that Rasch model-data fit statistics and monotonicity analyses may detect similar idiosyncratic rating patterns. It is also interesting to note that Rater 35 was identified for a violation of monotonicity, but this rater was classified as fair based on the Rasch model indices – suggesting that monotonicity analyses based on ac-MSA may be sensitive to rating patterns beyond what is detected by the Rasch model.

Table 3: Rating Quality Results

Rater	Rasch Rating Quality Indices			ac-MSA Rating Quality Indices		
	Measure	S.E.	Outfit *MSE*	Significant Violations of Monotonicity	Scalability	Significant Violations of IRO
1	0.37	0.37	**1.81**^{N}	1	0.12	27
2	0.02	0.02	0.92	0	0.34	6
3	-0.02	-0.02	1.10	0	0.31	8
4	0.26	0.26	0.71	0	0.30	17
5	0.61	0.61	1.02	0	0.32	6
6	0.35	0.35	1.12	0	0.29	11
7	**-0.57**^{L}	-0.57	1.09	0	0.22	6
8	0.09	0.09	1.11	0	0.29	10
9	0.31	0.31	1.09	0	0.31	10
10	**-0.65**^{L}	-0.65	**1.53**^{N}	1	0.30	5
11	0.45	0.45	1.37	0	0.23	15
12	**0.63**^{S}	0.63	0.70	0	0.29	11
13	0.25	0.25	0.78	0	0.27	16
14	**1.05**^{S}	1.05	0.67	0	0.28	5
15	**-0.71**^{L}	-0.71	0.94	0	0.32	4
16	**-1.50**^{L}	-1.50	0.70	0	0.22	1
17	**-0.86**^{L}	-0.86	**1.51**^{N}	0	0.20	10
18	-0.44	-0.44	0.74	0	0.23	7
19	**-0.64**^{L}	-0.64	0.55	0	0.19	7
20	0.20	0.20	0.95	0	0.18	21
21	**0.75**^{S}	0.75	**1.50**^{N}	0	0.17	20
22	0.13	0.13	1.16	0	0.28	11
23	0.31	0.31	1.01	0	0.28	7
24	**1.14**^{S}	1.14	0.80	0	0.30	2
25	-0.18	-0.18	0.87	0	0.26	5
26	**0.71**^{S}	0.71	1.02	0	0.32	6
27	**-0.77**^{L}	-0.77	1.37	0	0.21	9
28	**-0.96**^{L}	-0.96	**0.49**^{M}	0	0.21	2
29	0.18	0.18	0.83	0	0.31	11
30	0.10	0.10	1.04	0	0.28	5
31	0.06	0.06	0.71	1	0.26	12
32	0.39	0.39	0.92	0	0.36	6
33	**0.89**^{S}	0.89	1.17	0	0.28	6
34	-0.41	-0.41	0.68	0	0.23	8
35	-0.24	-0.24	0.74	1	0.37	4
36	**0.62**^{S}	0.62	0.79	0	0.36	10
37	**-0.86**^{L}	-0.86	0.91	0	0.16	2
38	**0.67**^{S}	0.67	0.71	0	0.38	10
39	**-1.14**^{L}	-1.14	**0.39**^{M}	0	0.18	1
40	**-0.60**^{L}	-0.60	1.41	0	0.20	10

Note. Superscripts are used to identify raters as follows: L = Lenient; S = Severe; M= Muted; N = Noisy

Table 4:
Average ac-MSA results within Rasch classifications

Rater Group	A-C Scalability		Significant Violations of Rater Monotonicity		Significant Violations of IRO	
	M	*SD*	*M*	*SD*	*M*	*SD*
Severe (*n*=8)	0.30	0.06	0.00	0.00	8.44	5.20
Lenient (*n*=11)	0.22	0.05	0.09	0.30	5.18	3.49
Noisy (*n*=4)	0.17	0.03	0.25	0.50	19.50	7.05
Muted (*n*=2)	0.20	0.02	0.00	0.00	1.50	0.71
Fair (*n*=20)	0.29	0.04	0.11	0.32	9.38	3.87

In addition to statistical tests for monotonicity, we also examined rater monotonicity plots within each of the rater error groups based on the Rasch model. Figure 3 includes the monotonicity plot for two randomly selected raters in each category. In these plots, student restscores are listed along the x-axis, where Restscore Group 1 has the lowest restscores, and Restscore Group 4 has the highest restscores. Examination of the monotonicity plots indicates differences in rating patterns across the groups. As may be expected, response functions for raters in the lenient group tend to have higher overall locations on the *y*-axis – indicating higher average ratings across all rest-score groups. Similarly, raters in the severe group tend to have lower overall response functions – indicating lower average ratings.

The response functions for raters in the noisy group indicate idiosyncratic use of the rating scale categories across levels of student achievement. In particular, these response functions are characterized by “dips” and “jumps” in the category probabilities across restscore groups that reflect violations of monotonicity. For example, Rater 1 displays haphazard rating patterns related to the second restscore group, and Rater 10 displays haphazard rating patterns related to the third restscore group. The plots for these raters also reveal category disordering within one or more restscore groups (i.e., the category probabilities are not ordered as expected based on the ordinal rating scale). The somewhat haphazard patterns suggest that these raters’ interpretations of student achievement were inconsistent with the other raters in the sample, whose ratings were used to classify students within rest-score groups. On the other hand, the response functions for raters in the muted group are generally steep and the distance between rating scale categories is less haphazard than the raters in the noisy group. Finally, the response functions for raters in the fair group are moderately steep, and are generally parallel and equidistant across the range of student rest-scores.

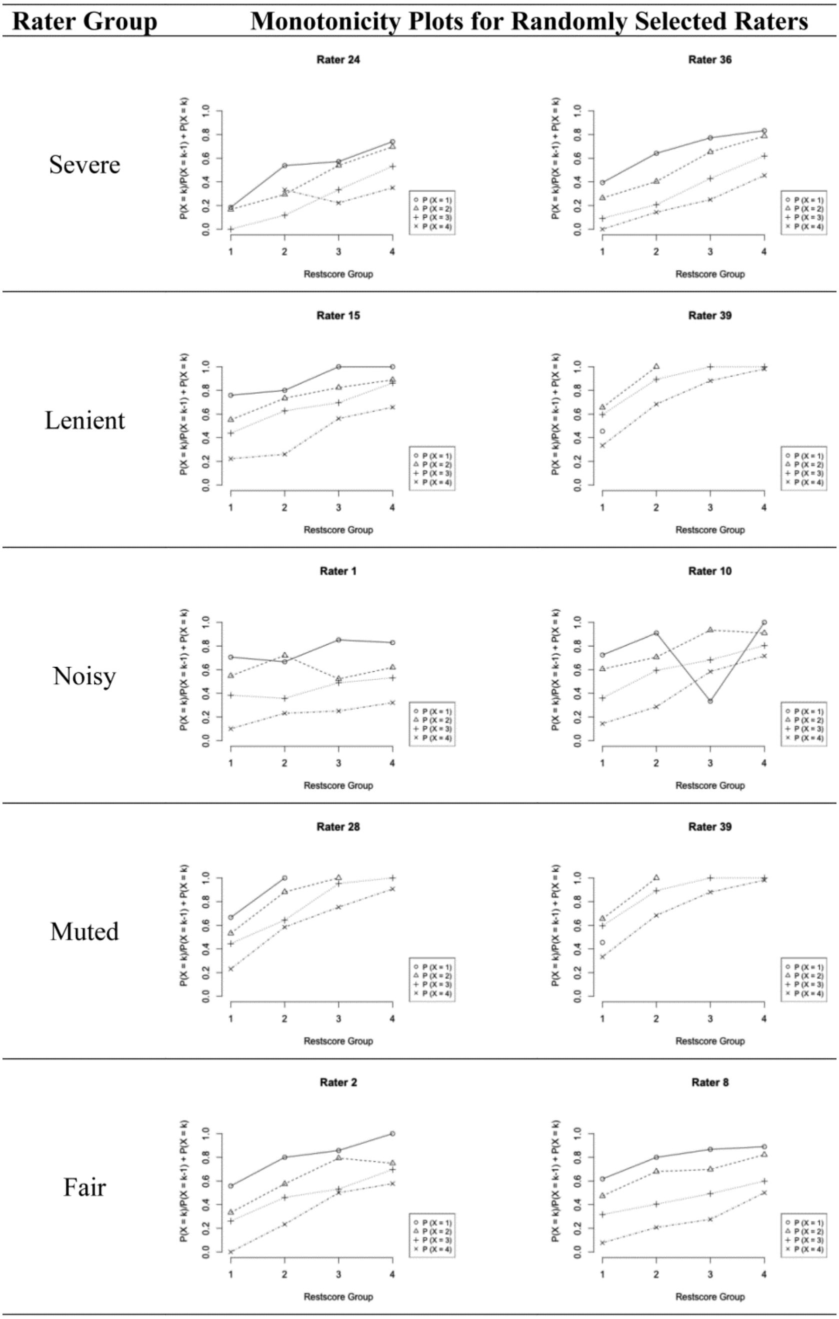


Figure 3:
Rater Monotonicity Plots within Rater Groups

B. Rater scalability

Table 3 includes rater scalability coefficients (H_i) for each of the 40 raters who scored the AHSW test. The coefficients range from 0.12 for Rater 1 to 0.38 for Rater 38, suggesting that Guttman errors were observed for each of the raters, and that there was some variation in the extent to which Guttman errors were observed across the group of raters.

When rater scalability is considered in terms of the Rasch classifications (Table 4), the lowest average scalability coefficients are observed within the noisy rater group (*M*=0.17, *SD*=0.03), followed by the muted rater group (*M*=0.20, *SD*=0.02). This finding suggests that Rasch model-data fit statistics and rater scalability coefficients based on ac-MSA may detect similar idiosyncratic rating patterns. It is interesting to note that the highest average rater scalability coefficients are observed within the severe rater group (*M*=0.20, *SD*=0.02) – suggesting that rater errors related to severity may not be associated with Guttman errors, as defined based on adjacent-categories probabilities.

C. Invariant rater ordering

Table 3 includes the frequency of significant violations of IRO for each of the 40 raters who scored the AHSW test. These results indicate at least one significant violation for each of the raters. The highest number of significant violations (*n*=27) was observed for Rater 1, followed by Rater 20 (*n*=21).

When IRO is considered in terms of the Rasch classifications (Table 4), it is interesting to note that violations occurred most frequently within the noisy rater group (*M*=19.50, *SD*=7.05), and least frequently within the muted rater group (*M*=1.50, *SD*=0.71). This finding suggests that IRO analyses based on ac-MSA may be sensitive to unexpected ratings that are also captured by Rasch fit statistics.

In addition to statistical tests for IRO, we also examined graphical displays of IRO using plots similar to the example shown in Figure 2. Overall, the graphical displays provided insight into not only whether a violation of IRO occurred, but also the nature of the violation. Of particular interest are the IRO plots for the raters who were most frequently associated with violations of IRO: Rater 1, who was classified as noisy and Rater 20, who was classified as fair. Selected IRO plots that represent the general patterns observed for these two raters are presented in Figure 4.

Across the IRO plots for Rater 1, who was classified as noisy, it was interesting to note that most of the significant violations of invariant ordering that involved Rater 1 occurred in conjunction with raters classified as fair. Inspection of the IRO plots in Figure 4 highlights the nature of these discrepancies in rater ordering across student achievement levels. Specifically, these plots reveal that the response function for Rater 1 (dashed line) is somewhat flat across student achievement levels, suggesting low discrimination. When Rater 1 was paired with raters who more clearly distinguished among levels of student achievement, Rater 1 was relatively more lenient for the lower achievement levels and relatively more severe for the higher achievement levels – resulting in a violation of invariant ordering.

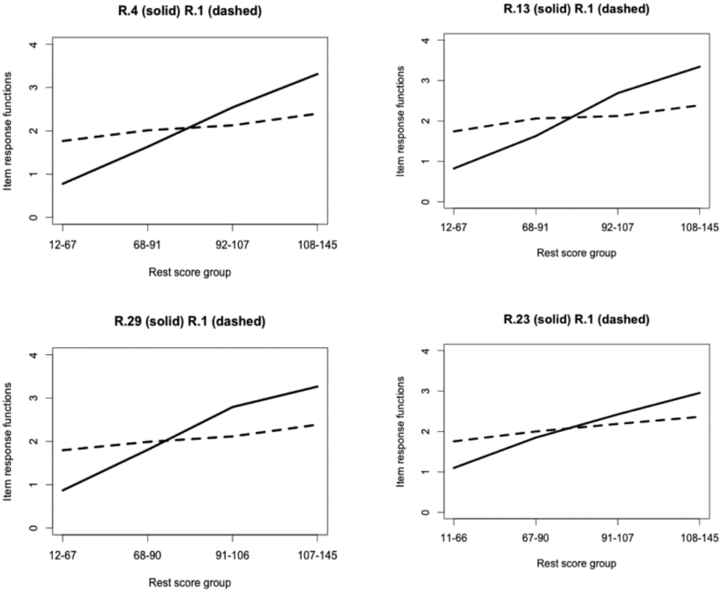
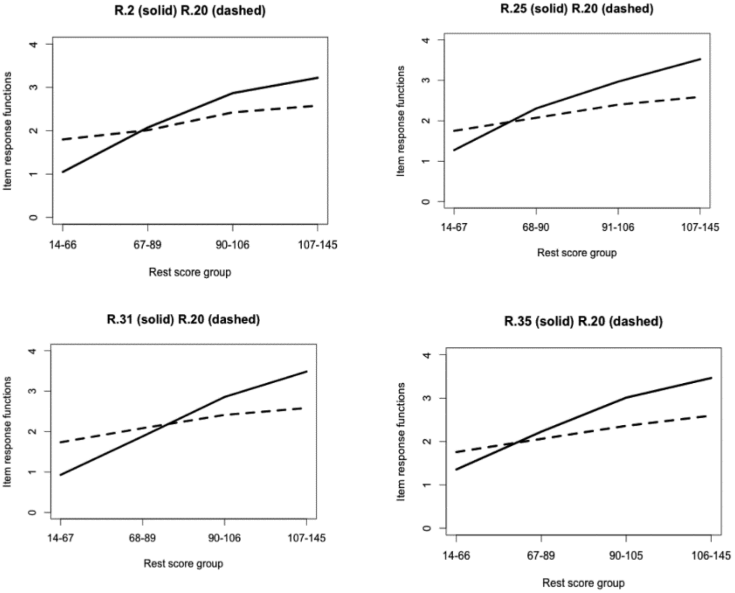
Rater	Selected Invariant Rater Ordering Plots
1	
20	

Figure 4:
Selected Invariant Rater Ordering (IRO) plots for the raters with the most frequent significant violations of IRO

Figure 4 also includes plots for Rater 20. Although Rater 20 was classified as fair based on the Rasch model, many violations of IRO were observed in association with this rater. Interestingly, inspection of the IRO plots for Rater 20 reveal a similar pattern as was observed for Rater 1. Specifically, the response function for Rater 20 (dashed line) is relatively flat across the restscore groups – suggesting low discrimination of student writing achievement compared to the other raters. As a result, combinations of this rater with raters with steeper slopes (higher discrimination) resulted in inconsistencies in the relative ordering of Rater 20 with other raters. Similar to Rater 1, most of the significant violations of IRO associated with Rater 20 involved other raters who were classified as fair.

Summary and Discussion

In this study, we explored a nonparametric method for examining the psychometric quality of rater-mediated assessments in terms of specific types of rater errors and systematic biases. Specifically, we used an adaptation of MSA, which is a nonparametric approach to IRT. MSA is a useful method for evaluating the psychometric properties of rater-mediated educational performance assessments because it allows researchers to explore the degree to which individual raters adhere to important measurement properties without parametric transformations. Accordingly, MSA can be used to explore the degree to which individual raters adhere to fundamental measurement properties, such as invariance, without the need for strict parametric model requirements.

In the current analysis, we used an adaptation of ac-MSA to explore the degree to which differences in rating quality indices based on MSA were observed within groups of raters who were identified as demonstrating specific types of rater errors and systematic biases using indicators from Rasch measurement theory. Overall, the results provided an overview of the correspondence between indicators of leniency/severity and response sets based on Rasch measurement theory and indicators of rater scalability, monotonicity, and invariant ordering based on ac-MSA.

We observed very few violations of rater monotonicity among the AHSW assessment raters in any of the rater groups. However, inspection of monotonicity plots for the groups of raters indicated differences in the application of rating scale categories across each of the rater groups. In particular, we observed that patterns in ISRFs not only revealed differences related to rater leniency/severity based on their overall locations, but also provided diagnostic information related to the application of rating scale categories across various levels of student achievement that reflected response sets. As noted by Wind (2016) and Wind and Engelhard (2015), one of the major benefits of the use of MSA is the utility of the graphical displays for providing diagnostic information related to the underlying measurement properties in a set of ratings.

In terms of scalability, the results suggest that violations of Guttman ordering were observed most frequently among raters who were classified as noisy. This finding is somewhat unsurprising, given the shared underlying theoretical relationship to Guttman scaling for both MSA and Rasch measurement theory (Engelhard, 2008; van Schuur, 2003).

The alignment illustrated in this study between rating patterns classified as “misfitting” for both MSA and Rasch measurement theory reflect this shared focus on the basic requirements for invariant measurement that also characterize Guttman scaling (Guttman, 1950).

When IRO was considered in terms of the Rasch classifications, it was interesting to note that similar patterns were observed for the noisiest rater in the sample and a rater who was classified as fair based on the Rasch model. This result suggests that indicators of rating quality based on ac-MSA provide insight into rating patterns that is not captured by Rasch model-data fit statistics. Furthermore, examination of the graphical displays for IRO revealed that significant violations of the invariant ordering requirement were frequently observed for raters whose response functions were relatively flat, regardless of their classification based on the Rasch model. This finding suggests that raters’ overall level of discrimination across achievement level groups contributes to the degree to which they contribute to an invariant ordering of rater severity across students.

Next, we return to the two guiding research questions for this study and discuss the major findings from our analysis related to each question. A discussion of the implications follows.

Research question One: How can MSA be used to detect rater leniency/severity?

Because MSA is nonparametric, it is not possible to calibrate raters on an interval-level scale, as in parametric models for raters. Instead, differences in rater severity are identified using average ratings across the range of student achievement (rest-scores). In this study we observed that rater monotonicity plots, which show nonparametric response functions for raters, revealed differences in overall rater severity. Results from the analyses also highlighted the diagnostic value of monotonicity plots for identifying specific ranges of student achievement within which differences in rater severity were most prevalent. Together, the current findings suggest that nonparametric indices of monotonicity can also be used to identify and explore differences in rater leniency/severity that go beyond overall calibrations and provide diagnostic information about differences in rating quality across the range of student achievement levels.

Research question Two: How can MSA be used to detect rater response sets?

Indicators of model-data fit to the Rasch model were used to identify raters whose use of the AHSW rating scale resulted in unexpected response patterns that were classified as either “noisy” or “muted.” Similar to the findings for rater leniency/severity, results from monotonicity analyses suggested that graphical indices of rater monotonicity can also be used to detect ranges of student achievement within which idiosyncratic application of rating scale categories are observed for individual raters. Further, results from rater scalability analyses revealed lower overall scalability coefficients among the group of "noisy" raters. This finding suggests that MSA indices of scalability for individual raters correspond to Rasch-based indicators of rater response sets, such that low values of rater scalability can be used to identify raters with idiosyncratic application of a rating scale.

Results from IRO analyses were also informative regarding rater response sets. In particular, our findings of large departures from the assumption of IRO among raters in the response set groups as well as raters who were classified as fair indicate that there was not a meaningful relationship between violations of IRO and raters' classification within the response set subgroups. Accordingly, these results suggest that ac-MSA highlights characteristics of rater judgment that is not captured by Rasch fit statistics.

Conclusions

Taken together, the results from this study suggest that differences in rater scalability, monotonicity, and invariant rater ordering reflect related but not equivalent characteristics of rating patterns as the rater errors and systematic biases that can be identified using indices based on the Rasch model. The shared theoretical and empirical underpinnings between Rasch measurement theory and ac-MSA related to invariant measurement are reflected in the correspondence between indices of rating quality between the two approaches (Engelhard, 2008; Meijer et al., 1990; van Schuur, 2003). On the other hand, the finding that the two approaches were not completely congruent suggests that ac-MSA can provide additional insight into the psychometric quality of ratings that is not captured by the more commonly used parametric approaches.

In terms of implications for practice, the results from this study suggest that rating quality indices based on ac-MSA can be used to identify rater severity/leniency and idiosyncratic rating patterns that warrant further investigation. These indices should be viewed as an additional set of methodological tools for exploring rating quality from the perspective of invariant measurement that complement the more-frequently used parametric indices based on Rasch measurement theory.

In terms of research, these findings have implications regarding the use of nonparametric methods for evaluating rating quality. The analyses in the current study build upon the initial explorations of MSA in general (Wind & Engelhard, 2015; Wind, 2017), as well as the use of ac-MSA models (Wind, 2016; Wind & Patil, 2016) in particular, as an approach for evaluating rating quality. In particular, the current findings provide a connection between ac-MSA indicators of monotonicity, scalability, and IRO and the rater errors and systematic biases that are frequently discussed in the literature on rating quality (e.g., Wind & Engelhard, 2012).

Language assessments are used around the world for various educational and occupational decisions. These assessments are typically scored by raters, and it is essential to evaluate the reliability, validity and fairness of the ratings based on their intended uses (AERA, APA, & NCME, 2014). This study illustrates a suite of indices based on a probabilistic-nonparametric approach that can be used to identify and diagnose potential rater errors and systematic biases without the application of a parametric model. As noted in other applications of nonparametric IRT, the nonparametric statistics and displays based on ac-MSA provide an exploratory approach to exploring data quality that highlight departures from important measurement properties, such as monotonicity, scalability and invariance. The current study highlighted the additional diagnostic benefit of statistics and displays based on ac-MSA for exploring leniency/severity and response sets within the context of rater-mediated assessments. Additional research is needed that explores the use of the ac-MSA indices of rater scalability, monotonicity, and invariant ordering to communicate information about rating quality to practitioners and raters during training and operational scoring.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. A. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*(2), 8–14. https://doi.org/10.1111/emip.12074

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main: Peter Lang.

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*(1), 19–33.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & G. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155–189. https://doi.org/10.1080/15366360802197792

Engelhard, G., & Wind, S. A. (in press). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Taylor & Francis.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, E. A. Suchman, P. F. Lazarsfeld, & S. A. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 60–90). Princeton, NJ: Princeton University Press.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1–9. https://doi.org/10.1016/j.asw.2007.05.002

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*(3), 331–347. https://doi.org/10.1007/BF02294555

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/ Praeger.

Ligtvoet, R., Van der Ark, L. A., Marvelde, J. M. te, & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578–595. https://doi.org/10.1177/0013164409355697

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2015). Facets Rasch measurement (Version 3.71.4). Chicago, IL: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298. https://doi.org/10.1177/014662169001400306

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85–110). New York, NY: Routledge.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, *3*(8), 145–164.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. https://doi.org/10.1016/j.paid.2010.08.016

Snijders, T. A. B. (2001). Two-level nonparametric scaling for dichotomous data. In A. Boomsa, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 319–338). New York, NY: Springer.

Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*(3), 273–282. https://doi.org/10.1177/01466210122032073

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163.

Wind, S. A. (under review). A weighted polytomous adjacent-categories scalability coefficient for Mokken scale analysis.

Wind, S. A. (2016). Adjacent-categories Mokken models for rater-mediated assessments. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164416643826

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12153

Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*(4), 321–335.

Wind, S. A., & Engelhard, G. (2015). Exploring rating quality in rater-mediated assessments Using Mokken scale analysis. *Educational and Psychological Measurement*, *76*(4), 685–706. https://doi.org/10.1177/0013164415604704

Wind, S. A., & Patil, Y. J. (2016). Exploring incomplete rating designs with Mokken scale analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644 16675393

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 026553221668699. https://doi.org/10.1177/0265532216686999

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x